What's Your Argument? A Detailed Investigation on Argument Detection and Understanding with LLMs

Anonymous ACL submission

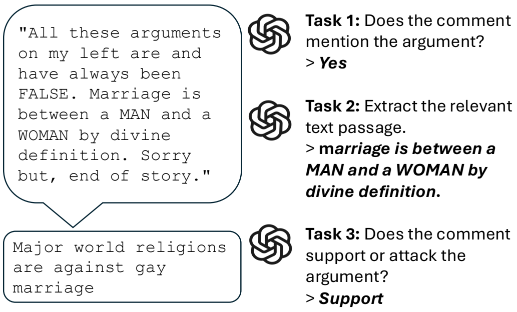
Abstract

Automated large-scale analysis of online argumentation around contested issues like abortion requires to detect and understand the usage of recurring arguments. Despite a large body of work in computational argumentation analysis, these tasks have not been tested with large language models. We fill this gap using a data set of over 2,000 opinion comments on polarizing topics and define three tasks: argument detection, extraction and identifying whether an argument is supported or attacked in a comment. We compare the performance of four state-of-the-art large language models (LLMs) and a fine-tuned RoBERTa baseline. We find that while LLMs excel at a binary support/attack decision, they can not reliably detect arguments in comments, and that performance does not consistently improve with in-context learning. We conclude by discussing the implications and limitations of current LLMs in argument-based opinion mining.

1 Introduction

Argumentation is the study of how humans express opinions, persuade others and reach conclusions, and is fundamental to human discourse and reasoning. In both formal and informal contexts, arguments form the basis of rational discourse, allowing individuals to present their viewpoints, support them with evidence, and engage in meaningful dialogue. The analysis of argumentative discourse has become increasingly critical in the digital age, where an unprecedented volume and velocity of online discourse shapes public opinion, policy decisions, and social movements (Lippi and Torroni, 2016).

This explosion of online discourse brings both challenges and opportunities for understanding human reasoning and opinion formation at scale. Automatic analysis of argumentative structures is crucial for tracking how opinions form and spread, identifying the evidence supporting different viewpoints, and evaluating the quality of public discourse (Stede and Schneider, 2018).



"All these arguments on my left are and have always been FALSE. Marriage is between a MAN and a WOMAN by divine definition. Sorry but, end of story."

Major world religions are against gay marriage

Task 1: Does the comment mention the argument?
> ***Yes***

Task 2: Extract the relevant text passage.
> ***marriage is between a MAN and a WOMAN by divine definition.***

Task 3: Does the comment support or attack the argument?
> ***Support***

Figure 1: An opinion piece (top, left) and pre-defined argument (bottom, left). We predict whether the opinion makes use of the argument (Task 1), where it mentions the argument (Task 2) and whether it supports or attacks the argument (Task 3).

Public discourse around contested issues — from abortion over immigration to climate change — is often dominated by recurring arguments that are repeated over and over by different parties. To automatically 1) identify these arguments; 2) detect them in the discourse and 3) understand how they are used (supported or attacked) would be an important contribution to automatic argument analysis. However, the majority of methods in opinion mining (Sun et al., 2017; Lawrence and Reed, 2015) and sentiment analysis (Bakliwal et al., 2013; Elghazaly et al., 2016; Ramteke et al., 2016) fall short of 2) and 3) by not identifying individual arguments and their way of use. Most work in argument mining, on the other hand, focuses on individual premises and claims without abstracting to broader cross-cutting arguments (Habernal and Gurevych, 2017; Lawrence and Reed, 2019). It remains unclear to what extent large language models (LLMs) can truly understand logical reasoning (Yan et al., 2024), or how well they grasp persuasive argumentation, which could potentially be used to spread misinformation and propaganda targeted at specific demographic groups (Rescala et al., 2024). This paper addresses this gap by testing state-of-the-art LLMs on a variety of argumentation tasks that would bring us closer to this goal.

Concretely, we here focus on the detection and way of usage of a set of pre-defined arguments in tasks 2, 3 – because we have to ensure that these tasks succeed before we can develop methods that identify arguments from the bottom up (Task 1). We leverage datasets comprising over 2,000 opinion comments covering six polarizing topics, from gay marriage to marijuana legalization (Boltužić and Šnajder, 2014; Hasan and Ng, 2014). Each topic is associated with a pre-defined set of established arguments, and each comment was manually annotated for the presence of each argument, and its usage (support or attack). Figure 1 (left) shows an example comment-argument pair, and Figure 2 shows excerpts of comments that support or attack an argument.

Given a pair of an opinionated text and an argument, we decompose our objective into three tasks (Figure 1, right) given a comment-argument pair: 1) predict whether the comment mentions the argument; 2) extract the span that expresses the argument; 3) identify whether the argument is *supported* or *attacked* in the comment.

We experiment with four state-of-the-art large language models (GPT4o, GPT4o-mini, Gemini1.5-flash and Llama3-8b). We find that a fine-tuned LLama model surpasses prompt-based LLMs, and that LLMs – especially the largest ones – outperform a fine-tuned RoBERTa baseline on argument detection and relationship identification, albeit barely so on identifying the spans of text that use an argument. Additionally, we find that in-context learning does not necessarily result in performance enhancement. We conclude with a discussion around implications, and important directions for future work to support accurate and reliable large-scale analysis of public discourse.

In sum, the contributions of this paper are:

- Investigating the ability of LLMs to understand and process argumentation, with a focus on identifying major cross-cutting arguments and their use in discourse.
- Evaluating four state-of-the-art LLMs across three argumentation tasks, highlighting that increasing the number of instruction examples does not always enhance performance, and demonstrating that small but fine-tuned models perform competitively with LLMs.
- Discussing the limitations, potential risks and ethical implications of LLMs in argumentation, offering directions for future work.

2 Related Work

Argument mining A vast body of work has studied the mechanisms of argumentation from theoretical and empirical points of view. Argument structure analysis starts with the identification of key argumentative elements, most typically premises and claims (Habernal and Gurevych, 2017; Hidey et al., 2017; Feng and Hirst, 2011). Claims present the speaker's position on a topic, while premises provide a justification for these claims (Hidey et al., 2017; Palau and Moens, 2009).

A second task involves determining how argument components interact with each other, with the goal to recognize whether a premise *attacks* or *supports* a claim (Cocarascu and Toni, 2017; Carstens and Toni, 2015; Ruiz-Dolz et al., 2021; Bench-Capon, 2003). Often argument detection and relation classification are performed jointly (Egawa et al., 2020; Stab and Gurevych, 2017).

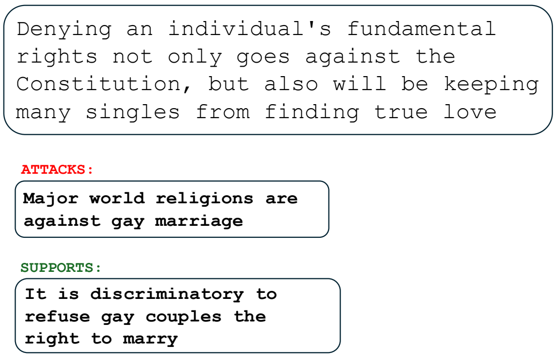
Argument structure analysis faces significant challenges, as the identification of claims is subjective, with no clear linguistic consensus on their precise definition or characteristics (Daxenberger et al., 2017), and is correspondingly hard to evaluate (Mestre et al., 2022). Furthermore, most work identifies arguments on an ad-hoc, document-level basis without mapping them back to broader recurring claim types which cross-cut the discourse making them less useful to map out patterns in broader discourse. We fill this gap by testing LLMs for identifying cross-cutting arguments and their relations and use pre-defined premises in this study to circumvent the challenge of evaluating model-identified claims.

Argument-based opinion analysis combines stance detection with argument structure into a framework for analyzing how people express their views (Arumugam, 2022). We build on early work which developed specialized corpora for studying argumentation by intersecting opinion-rich texts on divisive issues (like abortion) with pre-defined lists of related arguments (Boltužić and Šnajder, 2014; Hasan and Ng, 2014). These datasets provide comment-argument pairs manually labeled for their argumentative relationship (support or attack) (Boltužić and Šnajder, 2014), or manually highlighted the part of a comment that expresses a given argument (Hasan and Ng, 2014).

In a related line of work, *key point analysis* aims to identify standardized 'key points' in short 'arguments' with the goal of summarizing large collections of opinionated texts (Bar-Haim et al., 2020a) and has constructed large data sets of argument-key point pairs (Bar-Haim et al., 2020b, 2021; Tang et al., 2024). There are two fundamental differences to our work. First, their arguments (comparable to our 'comments') are (a) crowd-sourced from participants not necessarily invested or interested in the given issue, and (b) very short and hence less representative to actual argumentative text encountered in online discussions. The comments used in this work are drawn from online debate platforms. Secondly, key point analysis datasets allow to study argument *detection* but they do not reflect *how* arguments are used (support vs attack).Though our datasets are smaller, they enable a more comprehensive evaluation of models' argumentative capabilities, including identifying relevant spans and classifying attack/support relationships between comments and arguments.

Argument mining with Large Language Models Recently, LLMs have shown impressive performance in a variety of natural language tasks (Raiaan et al., 2024; Karanikolas et al., 2023), and argument mining is no exception. Recent works on argument pair extraction (de Wynter and Yuan, 2024), relation-based argument mining (Gorur et al., 2024; Otiefy and Alhamzeh, 2024), argument quality prediction (van der Meer et al., 2022) have shown performance gains with state-of-the-art LLMs. However, some other works have highlighted limited performance of LLMs in argumentation tasks, in particular in argument generation and persuasiveness (Hinton and Wagemans, 2023) and the identification of argumentative fallacies (Ruiz-Dolz and Lawrence, 2023). Other work has demonstrated a high ability of LLMs to detect persuasive arguments (Rescala et al., 2024) and evaluate argument quality (Mirzakhmedova et al., 2024).

The most comprehensive systematic review of LLMs performance in argument mining (AM) and argument generation tasks to date is Chen et al. (2024). The authors performed zero-shot and k-shot experiments using GPT-3.5, Flan-T5 family models as well as Llama2 models on a variety of AM tasks (claim detection, evidence detection, stance detection, evidence classification), as well as argument generation and summarization. Their results highlight decent performance on binary classification tasks (GPT3-5-Turbo and Flan-UL2 performing best in the zero-shot setup), but worse with more complex, multi-label classification tasks. Interestingly, a separate review focussed on the legal domain by Alsubhi et al. (2023) (including GPT-3.5 and GPT-4) concluded that these LLMs did not surpass their baseline (domain-specific BERT model). The same models also did not surpass the RoBERTa baseline in Ruiz-Dolz et al. (2024).


Denying an individual's fundamental rights not only goes against the Constitution, but also will be keeping many singles from finding true love

ATTACKS:

Major world religions are against gay marriage

SUPPORTS:

It is discriminatory to refuse gay couples the right to marry

Figure 2: Illustration of a comment supporting a pro-same-sex marriage argument and attacking a con-same-sex marriage argument.

This paper complements existing work as follows. First, we focus on identifying cross-cutting arguments within the context of a polarizing issue with LLMs. Given the contradicting results from previous surveys, we provide an important additional perspective. By formalizing three well-defined tasks we identify concrete shortcomings and formulate recommendations for future work in argumentation. Assessing the abilities of the LLMs to perform these tasks is of crucial importance to analyze opinions not just on a general level (such as stance, the argumentation structure, or key point summarization), but to support a nuanced analysis of the arguments on which such stance is based, allowing for greater scalability across topics.

3 Methodology

3.1 Data

Our study builds on prior research in natural language processing, particularly works that intersected curated arguments from online debate platforms with large-scale online argumentation data.

The **COMARG dataset**: Boltužić and Šnajder (2014) manually annotated 373 comments from the discussion platform *Procon.org* with a pre-defined list of arguments retrieved from *Idebate.org*. It encompasses two topics: gay marriage and the inclusion of the phrase "Under God" in the U.S. Pledge of Allegiance. The gay marriage-related comments were annotated for three arguments in favor and four arguments against the topic, while the Pledge of Allegiance topic featured three pro and three against arguments. Each comment-argument pair was further classified based on whether the comment supported, attacked, or made no use of the argument, as well as whether the support/attack was explicit or implicit. The inter-annotator agreement was moderate, and the final labels were decided by majority vote, excluding comment-argument pairs where no majority was reached.

The **YRU dataset**: Hasan and Ng (2014) sourced 1900 comments from an online debate platform, and their data set spans four topics: abortion, gay rights, legalization of marijuana, and the Obama presidency. For each topic, annotators identified a set of recurring arguments in the topics, leading to between 6 and 9 arguments each supporting and opposing the topic. The data set was originally developed for the task of argument extraction, i.e., identifying spans of text that employed a specific argument. Annotator agreement on this labelling task was reported as moderate to high, and disagreements were resolved through discussion.

All arguments for all six topics across the two data sets are listed in Table 4 in the Appendix.

3.2 Task Definitions

We define three argument mining (AM) tasks designed to test models' abilities to *detect* and *understand the use of* recurring arguments in collections of opinion texts.

Binary Argument Detection Given an argument A and a comment C, the task is to classify, in binary fashion, whether C makes use of A. We run this task on both the YRU anc COMARG data, across a total of six topics.

Argument Extraction Given an argument A and a comment C, the goal is to automatically detect the span within C that expresses A. Only the COMARG data set comes with manually-annotated argument spans, so we evaluate this task over the four COMARG topics.

Argument Relationship Classification Given an argument A and a comment C, we determine the relationship between A and C as C either attacking or supporting A (cf., Figure 2). We consider two formulations of this task: either a binary classification as support or attack; or a 4-way classification distinguishing between explicit/implicit support for or an explicit/implicit attack of an argument. Only the YRU data set labels the type of usage of an argument, so we evaluate relation classification over the two topics in this data set. A list of all pre-defined arguments from the original dataset is reported in the Appendix (Table 4).

3.3 Data Pre-Processing

Binary argument detection was conceptualized as a binary classification problem, necessitating preprocessing of the original data sets to conform to a binary format. For the COMARG dataset we consider all comment-argument pairs labeled as exhibiting any form of argumentative relationship as present (1). The data contained an explicit label of 'makes no use of an argument', which we reuse as our negative (not present) label (0). The YRU dataset is annotated for arguments on the sentence level. We project these labels to the comment-level, and consider them as present (1). All arguments not identified in any sentence were labeled as not present (0).

For Task 2, we treated the data in the COMARG dataset differently for the two subtasks. In subtask 2a we conflated the original labels in a binary fashion, only aiming at identifying whether the comment supports or attacks the argument. For subtask 2b we considered the original scale of implicit/explicit support and attack, we thus left the original labels unaltered. Finally, for Task 3 we only considered the labels present in the original YRU dataset and the manually annotated spans in the comment.

3.4 Models

We selected four Large Language Models (LLMs) from different model families, spanning both open-source and proprietary architectures: one open-source – Llama3-8b-Instruct (Dubey et al., 2024) – and three proprietary models; GPT4o-mini and GPT 4o (Achiam et al., 2023), as well as Gemini1.5-Flash (Reid et al., 2024). We followed established practices to minimize non-deterministic behavior and output variability (Zhang et al., 2023; Meng et al., 2023), i.e. setting the temperature to 0 and the top_p parameter to 1 (Liu et al., 2023; Brown et al., 2023). [1]

Prompts In preliminary experiments, we experimented with prompt variations along three key dimensions: structure (unstructured vs. structured step-by-step instructions), specificity (varying level of detail on task requirements and constraints), and role assignment (including/excluding the specific assignment of a role such as "you are an expert in argument analysis"). For argument detection (Task 1), a structured prompt with detailed instructions but without role assignment performed best. For both span extraction (Task 2) and argument relationship classification (Task 3), prompts that combined structured step-by-step instructions with explicit role assignment achieved superior performance. These optimized prompts were used for all subsequent experiments. The full prompts are in Appendix B (Tables 7 to 8).

Each task was attempted as zero-shot, 1-shot and 5-shot. To assess the impact of different examples, each few-shot experiment was run five times with randomly sampled, non-overlapping instruction examples to study the impact of chosen examples on the final results. We manually verified that examples were instructive, and that the five-shot example set covered all classes.

RoBERTa Baselines We fine-tuned one RoBERTa model (Liu, 2019) for each task, by combining all the available data across topics. The relatively small number of samples for individual topics renders topic-wise fine-tuning infeasible.

For the classification tasks, we concatenated each comment-argument pair using the [SEP] token as a delimiter. We randomly split the data into five stratified folds for cross-validation, ensuring a balanced label distribution in each split. Each model was trained for 3 epochs with a batch size of 16. For the span extraction task, we formatted the data equivalent to extractive question-answer tasks, where arguments serve as "question", and relevant spans as the "answer" to be extracted. We fine-tuned a RoBERTa model on this data using the QuestionAnsweringModel from SimpleTransformers[2] again with five fold stratified cross validation, training for a total of 10 epochs and with a batch size of 16.[3]

LLM Fine-tuning To disentangle the effect of fine-tuning from model size, we also fine-tune one of our LLMs. For Llama3-8b-Instruct we performed parameter-efficient fine-tuning using low-rank adaptation (LoRA) (Hu et al., 2021), with cross-validation on five stratified folds. The details of hyperparameters and training protocol are provided in Appendix D.We include fine-tuned Llama only for tasks 1 and 2 because for task 3 fine-tuned RoBERTa was widely outperformed by all LLMs in the prompting setup.

4 Results

We now present the quantitative results of our four LLMs and baselines across tasks. Overall, we find that (1) fine-tuned Llama achieves superior performance over all other models in detecting and extracting arguments; (2) larger LLMs generally outperformed smaller models and are more robust to different few-shot examples (exhibiting smaller variance); (3) that instruction examples (one- or five-shot) do not necessarily lead to enhanced performance; and (4) that the *detection* of arguments in comments (Task 1) is challenging for LLMs, which calls for caution with and future research on automated argument extraction.

4.1 Task 1: Binary Argument Detection

We test four models (Llama, GPT4o, GPT4o-mini, Gemini) in zero-, one-, and five-shot settings across six different topics on predicting whether a given argument is stated in a comment or not. Results in Table 1 show that all LLMs outperform the baselines, and that the fine-tuned Llama3 performs best overall [4]. Among the prompt-based models, the largest variants (GPT4o and Gemini) outperform their smaller counterparts. We observe a strong variance across topics, with abortion (AB) and gay marriage (GM) performing best. Finally, and perhaps counterintuitively, we do not observe consistent improvement with more examples. The standard deviation (std) across five model runs for few-shot experiments was ±0.01 to ±0.02 for larger

[1] For Llama3-8b-Instruct, we also set the top_k parameter to 1. GPT40-mini and Gemini1.5Flash do not feature manual configuration of this parameter.

[2] `https://simpletransformers.ai/docs/qa-model/`

[3] Information about the parameters are reported inc Appendix C.

[4] For task 1, the F1 SDs of the fine-tuned LLM range from ±0 to ±0.01, indicating robustness.

Model	GM			UG			AB			GR			MA			OBl			Combined		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority	0.33	0.50	0.40	0.34	0.50	0.41	0.45	0.50	0.47	0.44	0.50	0.47	0.43	0.50	0.46	0.46	0.50	0.48	0.40	0.50	0.44
RoBERTa																			0.67	0.60	0.61
	Zero shot																				
Gemini1.5-f	0.83	0.75	0.79	0.77	0.70	0.73	0.66	0.82	0.73	0.63	0.72	0.67	0.61	0.74	0.66	0.61	0.74	0.67	0.74	0.75	0.72
GPT4o	0.86	0.67	0.76	0.80	0.70	**0.75**	0.79	0.84	**0.81**	0.75	0.70	0.72	0.73	0.63	**0.68**	0.66	0.66	0.66	0.73	0.67	0.68
GPT4o-m	0.86	0.67	0.75	0.80	0.70	0.74	0.69	0.83	0.76	0.63	0.72	0.67	0.61	0.71	0.66	0.62	0.73	0.67	0.74	0.65	0.69
Llama3	0.69	0.68	0.69	0.65	0.66	0.65	0.59	0.72	0.65	0.61	0.71	0.65	0.57	0.70	0.63	0.59	0.68	0.63	0.66	0.63	0.65
	One shot																				
Gemini1.5-f	0.83	0.76	**0.80**	0.77	0.72	**0.75**	0.67	0.82	0.74	0.63	0.73	0.68	0.61	0.74	0.67	0.61	0.74	0.67	0.74	0.75	0.72
GPT4o	0.84	0.68	0.75	0.76	0.70	0.73	0.74	0.84	0.79	0.73	0.72	**0.73**	0.63	0.67	0.65	0.62	0.73	**0.68**	0.75	0.70	0.73
GPT4o-m	0.82	0.74	0.78	0.63	0.64	0.63	0.68	0.83	0.75	0.63	0.72	0.67	0.62	0.73	0.67	0.62	0.73	0.67	0.75	0.65	0.70
Llama3	0.63	0.63	0.63	0.62	0.64	0.63	0.59	0.66	0.62	0.61	0.56	0.63	0.57	0.61	0.59	0.59	0.61	0.60	0.62	0.60	0.61
	Five shot																				
Gemini1.5-f	0.83	0.77	**0.80**	0.76	0.73	0.74	0.66	0.82	0.73	0.62	0.72	0.67	0.61	0.74	0.67	0.61	0.74	0.67	0.74	0.73	0.73
GPT4o	0.84	0.69	0.76	0.75	0.69	0.72	0.70	0.84	0.76	0.70	0.73	0.71	0.64	0.69	0.66	0.65	0.71	**0.68**	0.73	0.67	0.71
GPT4o-m	0.78	0.72	0.75	0.63	0.64	0.63	0.68	0.83	0.75	0.63	0.73	0.68	0.63	0.74	**0.68**	0.62	0.74	0.67	0.73	0.65	0.70
Llama3	0.61	0.60	0.60	0.61	0.62	0.62	0.59	0.60	0.61	0.57	0.54	0.63	0.59	0.60	0.59	0.59	0.60	0.59	0.61	0.59	0.60
Llama3 FT																			0.77	0.74	**0.76**

Table 1: Results for binary argument detection (Task 1) for six topics and the combined data set (final column) as macro-averaged precision, recall and F1. We report a majority baseline, and fine-tuned RoBERTa and fine-tuned Llama3 (Llama3 FT) on the combined data only. The best F1 scores per data set are bolded. 1-shot and 5-shot results are averaged over five runs.

models, indicating high robustness to varying inputs, while smaller models showed higher std, typically ±0.02 to ±0.03, especially in 1-shot settings.

4.2 Task 2: Argument Extraction

Here, we tasked models with identifying the exact span of text in which an argument is being used. We report ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) between predicted and golden spans.

Results in Table 2 reveal that, similar as in Task 1, the fine-tuned Llama3 outperformed all other models [5]. In prompting experiments, 5-shot Gemini consistently performs best. We observe a consistent improvement with exposure to more examples in the task instruction. We posit that this is due to the extractive nature of the task, which is more challenging for LLMs out-of-the-box compared to the classification task (Task 1). Most interestingly, we observe that most LLMs outperform the RoBERTa baseline only in the 5-shot setting on the Combined data set, and the gap between non-fine tuned LLMs and RoBERTa is small (with the exception of 5-shot Gemini). For Task 2, larger models (Gemini, GPT4o) show low std (±0.01 to ±0.03), while smaller models (GPT4o-mini, Llama) exhibit slightly higher std (±0.02 to ±0.05), especially in 5-shot settings.

4.3 Task 3: Argument Relationship Classification

Given a comment and an argument featured in the comment, we ask models whether the argument is *supported* or *attacked* in the comment, either in a **binary** fashion, or on a 4-way **scale** (strongly/weakly supports; weakly/strongly attacks). Focusing on the binary task (Table 3, left) we observe that the two largest models (Gemini and GPT4o) consistently perform best, achieving almost perfect results. Exposure to examples does not improve performance and, in fact, substantially decreases results for GPT4-mini and Llama3. We observe a substantial performance decrease when moving to the 4-way classification task (Table 3, right), with the larger LLMs again performing best. The F1 std for the models show that Gemini1.5-f indicates low variability (std ±0.02), while GPT-4o-m and GPT-4o have substancial variability (std ±0.03 to ±0.16), and Llama3 shows even higher variability (std ±0.07 to ±0.10).

Interestingly, performance across models was higher in the binary version of Task 3 than Task 1. In other words, models do better at identifying

[5] With F1 standard deviations ranging from 0.01 to 0.015 across the folds, indicating stability

Model	AB			GR			MA			OB			Combined		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
RoBERTa													0.45	0.41	0.44
	Zero shot														
Gemini1.5-flash	0.42	0.41	0.42	0.41	0.40	0.41	0.37	0.36	0.37	0.38	0.36	0.38	0.40	0.38	0.40
GPT4o	0.31	0.30	0.31	0.32	0.31	0.32	0.30	0.28	0.30	0.32	0.30	0.32	0.31	0.30	0.31
GPT4o-m	0.28	0.27	0.28	0.29	0.27	0.29	0.27	0.25	0.27	0.25	0.23	0.25	0.27	0.26	0.27
Llama3	0.29	0.27	0.29	0.33	0.31	0.33	0.28	0.26	0.27	0.28	0.27	0.28	0.30	0.28	0.29
	One shot														
Gemini1.5-flash	0.46	0.45	0.46	0.46	0.45	0.46	0.43	0.41	0.43	0.47	0.46	0.47	0.46	0.44	0.46
GPT4o	0.36	0.35	0.36	0.41	0.39	0.41	0.37	0.36	0.37	0.41	0.39	0.41	0.39	0.37	0.39
GPT4o-m	0.35	0.34	0.35	0.38	0.36	0.38	0.37	0.35	0.37	0.36	0.35	0.36	0.37	0.35	0.37
Llama3	0.36	0.35	0.36	0.42	0.41	0.42	0.37	0.36	0.37	0.41	0.40	0.41	0.39	0.38	0.39
	Five shot														
Gemini1.5-flash	0.50	0.49	**0.50**	0.51	0.50	**0.51**	0.48	0.46	**0.48**	0.56	0.54	**0.55**	0.51	0.50	0.51
GPT4o	0.44	0.43	0.44	0.48	0.47	0.48	0.42	0.41	0.42	0.47	0.46	0.47	0.45	0.44	0.45
GPT4o-m	0.43	0.42	0.43	0.47	0.45	0.46	0.42	0.41	0.42	0.43	0.42	0.43	0.44	0.43	0.44
Llama3	0.48	0.47	0.48	0.50	0.49	0.50	0.43	0.41	0.43	0.50	0.48	0.50	0.48	0.46	0.48
Llama3 FT													0.55	0.50	**0.54**

Table 2: Results for Argument Extraction (Task 2) for the four topics in the YRU data set and the combined data set (final column) as Rouge 1, 2 and L. Models as in Table 1. The best Rouge-L scores per data set are bolded. 1-shot and 5-shot results are averaged over five runs with different examples.

whether a comment *supports or attacks* a given argument than at detecting whether a comment *uses* the argument. The models benefited from examples uniformly only for argument extraction (Task 2), but not in the classification tasks. Consistently, a fine-tuned RoBERTa model performed competitively with the LLMs on Task 2. Overall, we conclude that there is substantial room for improvement in LLM argument detection and interpretation for all presented task with the exception of binary argument relation classification.

4.4 Exploratory Analysis

A natural question following from the results above is where exactly LLMs fail on fine-grained argument detection and interpretation. As a step towards answering this question we conducted an exploratory analysis on argument detection (Task 1), which is the most comprehensive in terms of samples, and which revealed substantial room for improvement. We inspected the results in Table 1 by argument type (arguments in favor or against an issues), taking into consideration the prevalence of arguments in the golden data (determined as the frequency of an argument divided by the total number of arguments in the topic)[6]

Our analysis shows a clear trend of arguments with higher proportions within a topic tend to achieve higher F1 scores (a linear regression model showed a significant effect and R^2 of 0.26). We posit that arguments that are prevalent in our gold data are also more frequently discussed in general, leading to more exposure in the LLM training data and hence a better understanding. The two most frequent and most well-predicted arguments are "*Separation of state and religion*" (against UGIP; Proportion = 0.385, F1 = 0.76) and "*Gay people should have the same rights as straight people*" (pro GM; Proportion = 0.317, F1 = 0.72).

However, some interesting outliers challenge this trend. For example, we observe that some arguments with low proportions achieve relatively high F1 scores – e.g., "*Rape victims need it to be legal*" (pro abortion; Proportion = 0.057, F1 = 0.69) and "*Abortion should be allowed when a mother's life is in danger*" (pro abortion; Proportion = 0.042, F1 = 0.65). Both arguments are presented in relatively simple language, easing classification. Conversely, some relatively high-proportion arguments achieve low F1 scores. For example, "*Gay marriage undermines the institution of marriage, leading to an increase in out of wedlock births and divorce rates*" (Against GM; Proportion = 0.153, F1 = 0.12) is relatively frequent in the data set, but presumably challenging to classify due to its relatively higher complexity. We did not find any significant effect of the direction of arguments (pro vs against).

[6] Detailed information can be found in Appendix E.

Model	GM - binary			UG- binary			Comb- binary			GM - scale			UG - scale			Comb - scale		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority	0.31	0.50	0.39	0.29	0.50	0.37	0.30	0.50	0.38	0.10	0.25	0.14	0.29	0.50	0.37	0.19	0.37	0.25
RoBERTa							0.31	0.50	0.39							0.22	0.25	0.15
	Zero shot																	
Gemini1.5-f	0.91	0.93	0.92	0.96	0.96	**0.96**	0.93	0.94	0.94	0.55	0.56	0.55	0.58	0.60	0.59	0.56	0.58	0.57
GPT4o	0.93	0.95	0.94	0.95	0.97	**0.96**	0.94	0.96	**0.95**	0.52	0.57	0.56	0.59	0.63	**0.61**	0.55	0.60	0.58
GPT4o-m	0.77	0.77	0.77	0.90	0.91	0.91	0.83	0.84	0.84	0.41	0.39	0.40	0.35	0.46	0.40	0.38	0.42	0.40
Llama3	0.82	0.84	0.83	0.79	0.77	0.78	0.80	0.80	0.80	0.39	0.30	0.34	0.44	0.46	0.45	0.41	0.38	0.39
	One shot																	
Gemini1.5-f	0.91	0.94	**0.93**	0.89	0.90	0.90	0.90	0.92	0.91	0.56	0.58	**0.57**	0.60	0.62	**0.61**	0.58	0.60	**0.59**
GPT4o	0.73	0.70	0.71	0.86	0.87	0.86	0.80	0.78	0.78	0.41	0.39	0.40	0.35	0.47	0.40	0.38	0.43	0.40
GPT4o-m	0.66	0.63	0.65	0.81	0.81	0.81	0.73	0.72	0.73	0.38	0.36	0.37	0.33	0.44	0.38	0.35	0.4	0.37
Llama3	0.55	0.54	0.55	0.74	0.72	0.73	0.65	0.63	0.64	0.33	0.28	0.30	0.33	0.28	0.30	0.33	0.28	0.30
	Five shot																	
Gemini1.5-f	0.92	0.94	**0.93**	0.96	0.96	**0.96**	0.94	0.95	0.94	0.56	0.58	**0.57**	0.60	0.62	**0.61**	0.58	0.60	**0.59**
GPT4o	0.70	0.67	0.68	0.92	0.93	0.92	0.81	0.80	0.80	0.41	0.39	0.40	0.35	0.47	0.40	0.38	0.43	0.40
GPT4o-m	0.65	0.62	0.64	0.85	0.86	0.86	0.75	0.74	0.75	0.39	0.36	0.37	0.31	0.44	0.37	0.35	0.40	0.37
Llama3	0.54	0.54	0.54	0.75	0.72	0.74	0.64	0.63	0.64	0.30	0.27	0.29	0.30	0.27	0.29	0.30	0.27	0.29

Table 3: Results for Argument Relationship Classification (Task 3) for the two topics in the COMARG data set and the combined data set (final column) as macro precision, recall and F1. Left: binary classification (support vs attack); Right: 4-way classification (explicit/implicit support/attack). We compare against a majority baseline and fine-tuned RoBERTa model (combined data only). The best F1 scores per data set are bolded. 1-shot and 5-shot results are averaged over five runs with different examples.

5 Conclusion

We have presented a detailed investigation of how well LLMs can predict the presence of arguments in a text, the detection of the exact span in which is it present, and whether the comment supports or attacks the argument. Using a controlled testbed of texts with pre-defined arguments commonly discussed in polarizing topics, we focused on a nuanced evaluation of LLMs' understanding of argumentation, which is essential for large-scale argumentation analysis tasks.

While models excel at classification tasks (Task 1 and 3), their ability to extract specific argument spans (Task 2) is less convincing. And specifically for Task 2, a fine-tuned RoBERTa baseline was competitive. We fine-tuned Llama to disentangle the effect of fine-tuning from model size and found that fine-tuning massively improved the LLM compared to a prompt-based approach in zero- or few-shot settings. We note, however, that fine-tuning LLMs comes at a significant environmental and monetary cost and the practical value in the face of changing topics and arguments is questionable.

Contrary to expectations, the argument detection task was generally more challenging for models than the binary relationship identification, indicating that the models understand better how a comment uses an argument than whether the argument is present. We also found that one-shot or few-shot learning did not consistently improve model performance, with variation depending on the task, topic, and model. However, LLMs were robust to the selection of examples in the prompt as shown through largely low standard deviations across five runs.

Our exploratory analysis showed that arguments with higher proportions within a topic typically achieved higher F1 scores, but some low-proportion arguments with simpler language also performed well. Conversely, complex and ambiguous arguments posed challenges for the model. This suggests that both frequency and complexity of arguments impact argument detection and interpretation.

In conclusion, while LLMs perform well on traditional argumentation tasks, they are sensitive to argument frequency and complexity. Relying solely on LLM prompting techniques for argumentation analysis could lead to inaccurate classifications. Future work should explore how weaknesses can be addressed through improved prompting and fine-tuning, and further analyze the causes of performance disparities across different argument classes.

6 Limitation

The data used in this study is limited in scope, both in terms of size and the range of topics and arguments it covers. While this controlled data set enabled a detailed analysis of Large Language Models (LLMs) in argumentation tasks, it may not fully represent the complexity and diversity of real-world argumentative discourse. Notably, the datasets employed were released in 2014, and may not capture more recent arguments or shifts in public opinion. For instance, the arguments related to the subtopic of gay marriage might no longer be relevant, especially given the legalization of gay marriage in the US in 2015, shortly after the data was released. On account of the limited data set size, we needed to conflate all datapoints for Task 1 to fine-tune our RoBERTa baseline. Due to time and cost constraints, as well as environmental considerations, we were only able to fine-tune one LLM (Llama3) on the tasks.

7 Ethical Considerations

This study investigates the performance of LLMs in AM-related tasks on polarizing topics, which may involve sensitive or controversial discussions. We emphasize that the findings and conclusions of this research are not intended to amplify or legitimize harmful, discriminatory, or unethical viewpoints. Instead, the goal is to evaluate and enhance the understanding of LLMs' capabilities in argument detection, classification and extraction. Our research does not seek to endorse divisive or harmful opinions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sarah M. Alsubhi, Areej M. Alhothali, and Amal A. Al-Mansour. 2023. AraBig5: The Big Five Personality Traits Prediction Using Machine Learning Algorithm on Arabic Tweets. *IEEE Access*, 11:112526–112534.

S. S. Arumugam. 2022. Development of argument based opinion mining model with sentimental data analysis from twitter content. *Concurrency and Computation: Practice and Experience*, 34(15):e6956.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. *Preprint*, arXiv:2005.01619.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key point analysis of business reviews. *Preprint*, arXiv:2106.06758.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics.

T. J. M. Bench-Capon. 2003. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Filip Boltužić and Jan Šnajder. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Sarah Brown, Peter Anderson, and David Miller. 2023. Understanding the role of sampling parameters in language model generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3456–3470.

Lucas Carstens and Francesca Toni. 2015. Towards relation based Argumentation Mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2024. Exploring the Potential of Large Language Models in Computational Argumentation. *arXiv preprint*. ArXiv:2311.09022 [cs].

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Adrian de Wynter and Tangming Yuan. 2024. "I'd Like to Have an Argument, Please": Argumentative Reasoning in Large Language Models. In *Computational Models of Argument*, pages 73–84. IOS Press.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2020. Corpus for Modeling User Interactions in Online Persuasive Discussions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1135–1141, Marseille, France. European Language Resources Association.

Tarek Elghazaly, Amal Mahmoud, and Hesham A Hefny. 2016. Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of things and Cloud Computing*, pages 1–5.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can Large Language Models perform Relation-based Argument Mining? *arXiv preprint*. ArXiv:2402.11243 [cs].

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Martin Hinton and Jean HM Wagemans. 2023. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator. *Argument & Computation*, 14(1):59–74.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2023. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290.

John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Yanting Liu, Xue Zhang, and Brian Thompson. 2023. An empirical study of temperature parameter impact on large language model outputs. *Transactions of the Association for Computational Linguistics*, 11:845–862.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaolong Meng, Jianxin Wu, and Kai Chen. 2023. Enhancing reproducibility in large language models: A study of temperature and top-p parameters. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1123–1135.

Rafael Mestre, Matt Ryan, Stuart E Middleton, Richard Gomer, Masood Gheasi, Jiatong Zhu, and Timothy J Norman. 2022. Benchmark evaluation for tasks with highly subjective crowdsourced annotations: Case study in argument mining of political debates.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? *ArXiv*, abs/2404.09696.

Yasser Otiefy and Alaa Alhamzeh. 2024. Exploring Large Language Models in Financial Argument Relation Identification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 119–129, Torino, Italia. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *International Conference on Artificial Intelligence and Law*.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. 2016. Election result prediction using twitter sentiment analysis. In *2016 international conference on inventive computation technologies (ICICT)*, volume 1, pages 1–5. IEEE.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? *ArXiv*, abs/2404.00750.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barbera, and Ana Garcia-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

Ramon Ruiz-Dolz, John Lawrence, Ella Schad, and Chris Reed. 2024. Overview of DialAM-2024: Argument Mining in Natural Language Dialogues. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 83–92, Bangkok, Thailand. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Manfred Stede and Jodi Schneider. 2018. *Argumentation mining*. Springer.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25.

An Tang, Xiuzhen Zhang, and Minh Dinh. 2024. Aspect-based key point analysis for quantitative summarization of reviews. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1419–1433, St. Julian's, Malta. Association for Computational Linguistics.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31. IOS Press.

Junbing Yan, Chengyu Wang, Junyuan Huang, and Wei Zhang. 2024. Do large language models understand logic or just mimick context? *ArXiv*, abs/2402.12091.

Mei Zhang, Wei Chen, Yixuan Wang, and Hongzhi Li. 2023. Investigating the impact of decoding strategies on large language model performance: A systematic analysis. *arXiv preprint arXiv:2306.09265*.

A Lists of Arguments

Here, we present the complete list of pro and con arguments from the original datasets in Table 4.

B Prompts

We display the prompts used for our three tasks in Table 7 to Table 8.

C RoBERTa Fine-Tuning

We fine-tuned RoBERTa-base using the following configurations for each task:

- **Task 1: Argument Detection**
  - Training batch size: 16
  - Evaluation batch size: 64
  - Number of epochs: 3
  - Warmup steps: 500
  - Weight decay: 0.01
  - Evaluation strategy: per epoch
  - Save strategy: per epoch
  - Load best model at end: Yes
- **Task 2: Argument Extraction**
  - Training batch size: 16
  - Evaluation batch size: 16
  - Number of epochs: 10
  - Maximum sequence length: 512
  - N-best size: 16
  - Evaluate during training: No
  - Save checkpoints: No
  - Overwrite output directory: Yes
  - Save model every epoch: No
- **Task 3: Relationship Classification**
  - Training batch size: 16
  - Evaluation batch size: 64
  - Number of epochs: 3

- Warmup steps: 500
- Weight decay: 0.01
- Evaluation strategy: per epoch
- Save strategy: per epoch
- Load best model at end: Yes
- Optimization metric: F1
- Optimization goal: maximize

All models were trained on a single NVIDIA V100 GPU using the RoBERTa-base checkpoint as the initial model.

D Parameter-efficient finetuning (PEFT) of LlaMA

For PEFT, we used an implementation of low-rank adaptation (LoRA) from Unsloth AI[7] with the following hyperparameters:

- load in 4 bit = False
- r = 16
- target modules = q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
- lora alpha = 16
- lora dropout = 0
- bias = none
- use gradient checkpointing = unsloth
- use rslora (rank stabilized LoRA) = False

The finetuning was performed with 5-fold cross-validation (data split of 60-20-20 for train-dev-test sets, with test splits covering the whole dataset). For the classification task, the splits were stratified. The training used 8-bit Adam as optimizer and the standard learning rate of 2e-4. The number of training steps was proportional to the data size, with loss falling to near-zero values as a stop signal, and roughly amounted to 3 full epochs for the classification task and 5 full epochs for the span extraction task.

The same prompts and example/label formats were used for finetuning as for the zero-shot and few-shot experiments (see Appendix B).

E Detailed Results

Additionally, Table 9 to Table 14 report the full metrics for each subtopic for the per-argument analysis for the best-performing model in Task 1, as explained in Section 4.4. To better understand the relationship between argument proportions and model performance, we plotted the proportion of each argument within its topic against its corresponding F1 score, as shown in Figure 3. Each point represents an argument, with its proportion on the x-axis and its F1 score on the y-axis. The points are colored based on their stance, with red representing arguments against the issue ("CON") and blue representing arguments in favor of the issue ("PRO"). We also fitted a linear regression model (ordinary least squares) to assess the relationship between the proportion of argument in a topic and the argument F1 score. The model explained 26.2% of the variance ($R^2 = 0.262$) and showed a significant positive association (coefficient = 1.0758, $p < 0.001$), indicating that higher argument proportions predict higher F1 scores, as reported in Table 15.

[7] `https://github.com/unslothai/unsloth`

Data set	Pro Arguments	Con Arguments
GM	It is discriminatory to refuse gay couples the right to marry. Gay couples should be able to take advantage of the fiscal and legal benefits of marriage. Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology. Others	Gay couples can declare their union without resort to marriage. Gay marriage undermines the institution of marriage, leading to an increase in out-of-wedlock births and divorce rates. Major world religions are against gay marriages. Marriage should be between a man and a woman. Others
UG	Likely to be seen as a state-sanctioned condemnation of religion. The principles of democracy regulate that the wishes of American Christians, who are a majority, are honored. "Under God" is part of the American tradition and history. America is based on democracy and the pledge should reflect the belief of the American majority Others	Implies ultimate power on the part of the state. Removing "under God" would promote religious tolerance. Separation of state and religion. Others
AB	Abortion is a woman's right. Rape victims need it to be legal. A fetus is not a human yet, so it's okay to abort. Abortion should be allowed when a mother's life is in danger. Unwanted babies are ill-treated by parents and/or not always adopted. Birth control fails at times, and abortion is one way to deal with it. Abortion is not murder. Mother is not healthy/financially solvent. Others	Put the baby up for adoption. Abortion kills a life. An unborn baby is a human and has the right to live. Be willing to have the baby if you have sex. Abortion is harmful to women. Others
GR	Gay marriage is like any other marriage. Gay people should have the same rights as straight people. Gay parents can adopt and ensure a happy life for a baby. People are born gay. Religion should not be used against gay rights. Others	Religion does not permit gay marriages. Gay marriages are not normal/against nature. Gay parents cannot raise kids properly. Gay people have problems and create social issues. Others
MA	Not addictive. Used as a medicine for its positive effects. Legalized marijuana can be controlled and regulated by the government. Prohibition violates human rights. Does not cause any damage to our bodies. Others	Damages our bodies. Responsible for brain damage. If legalized, people will use marijuana and other drugs more. Causes crime. Highly addictive. Others
OB	Fixed the economy. Ending the wars. Better than the Republican candidates. Makes good decisions/policies. Has qualities of a good leader. Ensured better healthcare. Executed effective foreign policies. Created more jobs. Others	Destroyed our economy. Wars are still ongoing. Unemployment rate is high. Healthcare bill is a failure. Poor decision-maker. We have better Republicans than Obama. Not eligible as a leader. Others

Table 4: Pro and Con Arguments for All Subtopics and Data Sets

Analyze whether the following comment about {topic} contains a specific argument.
Argument to check for: {argument}
Instructions:
1. Determine if the comment explicitly or implicitly uses the given argument
2. Assign a binary label:
- 1 if the argument is present
- 0 if the argument is not present

Requirements:
- Only use 1 or 0 as labels
- Provide output in valid JSON format
- Do not repeat or include the input text in the response
- Focus solely on the presence/absence of the specific argument

Return your analysis in this exact JSON format:

```
    "id": "id", "label": label_value
```

Analyze the following comment in relation to the given argument:
where label_value must be either 1 or 0 (without quotes)
Comment to analyze:

Table 5: Prompt for Task 1

Task: Binary Classification of Arguments about {topic}
Input Text: {comment_text}
Target Argument: {argument_text}
Role: You are an expert in argument analysis and logical reasoning,
specializing in identifying rhetorical patterns in social discourse.
Step-by-Step Instructions:
1. Read the input text thoroughly
2. Evaluate the text's relationship to the target argument, examining:
- Direct support or opposition
- Implicit agreement or disagreement

3. Make a binary classification decision
4. Format the output according to specifications

Classification Rules:
- Label = 5: Comment supports/agrees with argument
- Label = 1: Comment attacks/disagrees with argument

Critical Requirements:
- Use ONLY specified labels (1 or 5)
- Do NOT quote or repeat input texts
- Return VALID JSON only

Output Schema: { "id": "{id}", "label": label_value must be 1 or 5 without quotes }
Input Text:

Table 6: Prompt for Task 2 - Binary

Task: Classification of Arguments about {topic}
Input Text: {comment_text}
Target Argument: {argument_text}
Role: You are an expert in argument analysis and logical reasoning,
specializing in identifying rhetorical patterns in social discourse.
Step-by-Step Instructions:
1. Read the input text thoroughly
2. Evaluate the text's relationship to the target argument, examining:
- Direct support or opposition
- Implicit agreement or disagreement
3. Make a binary classification decision
4. Format the output according to specifications
Classification Rules:
- Label = 5: Comment supports/agrees with argument
- Label = 4: Comment supports/agrees with argument implicitly/indirectly
- Label = 2: Comment attacks/disagrees with argument implicitly/indirectly
- Label = 1: Comment attacks/disagrees with argument
Critical Requirements:
- Use ONLY specified labels (1 or 5)
- Do NOT quote or repeat input texts
- Return VALID JSON only
Output Schema: { "id": "{id}", "label": label_value must be 1, 2, 4 or 5 without quotes }
Input Text:

Table 7: Prompt for Task 2 - Full Scale

Task: Text Span Identification for Arguments about {topic}
Target Argument: {argument_text}
Role: You are an expert in argument analysis and logical reasoning,
specializing in identifying rhetorical patterns in social discourse.
Step-by-Step Instructions:
1. Read the input text carefully
2. Locate exact text spans that:
- Directly reference the target argument
- Express the same idea as the argument
3. Extract the precise text span
4. Format the output according to specifications
Critical Requirements:
- Extract EXACT text only (no paraphrasing)
- Include COMPLETE relevant phrases
- Use MINIMUM necessary context
- Maintain ORIGINAL formatting
- Return VALID JSON only
Output Schema:
{ "id": "{id}",
"span": "exact_text_from_comment" must be verbatim quote
}
Input Text:

Table 8: Prompt for Task 3

Argument	F1	Stance	Support	Proportion (in topic)
It is discriminatory to refuse gay couples the right to marry	0.71	PRO	162	0.13
Major world religions are against gay marriages	0.63	CON	162	0.13
Marriage should be between a man and a woman	0.62	CON	180	0.14
Gay couples can declare their union without resort to marriage	0.57	CON	195	0.15
Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology	0.47	PRO	194	0.15
Gay couples should be able to take advantage of the fiscal and legal benefits of marriage	0.44	PRO	195	0.15
Gay marriage undermines the institution of marriage, leading to an increase in out of wedlock births and divorce rates	0.12	CON	197	0.15

Table 9: Average F1 scores, Stance, Support (total counts), and Proportion (in topic) for each argument across all splits and models, GM - Task 1

Argument	F1	Stance	Support	Proportion (in topic)
Separation of state and religion	0.76	CON	124	0.39
Under God is part of American tradition and history	0.67	PRO	92	0.29
Removing under god would promote religious tolerance	0.59	CON	43	0.13
America is based on democracy and the pledge should reflect the belief of the American majority	0.29	PRO	58	0.18
Implies ultimate power on the part of the state	0.23	CON	1	0.00
Likely to be seen as a state sanctioned condemnation of religion	0.10	PRO	4	0.01

Table 10: Average F1 scores, Stance, Support (total counts), and Proportion (in topic) for each argument across all splits and models, UGIP - Task 1

Argument	F1	Stance	Support	Proportion (in topic)
Abortion is a woman's right	0.70	PRO	107	0.15
Rape victims need it to be legal	0.69	PRO	40	0.06
A fetus is not a human yet, so it's okay to abort	0.68	PRO	130	0.19
Abortion should be allowed when a mother's life is in danger	0.65	PRO	30	0.04
Abortion kills a life	0.63	CON	106	0.15
Be willing to have the baby if you have sex	0.63	CON	50	0.07
Unwanted babies are ill-treated by parents and/or not always adopted	0.60	PRO	38	0.05
An unborn baby is a human and has the right to live	0.60	CON	98	0.14
Birth control fails at times and abortion is one way to deal with it	0.37	PRO	12	0.02
Abortion is harmful for women	0.35	CON	11	0.02
Mother is not healthy/financially solvent	0.29	PRO	21	0.03
Abortion is not murder	0.23	PRO	18	0.03
Put baby up for adoption	0.12	CON	38	0.05

Table 11: Average F1 scores, Stance, Support (total counts), and Proportion (in topic) for each argument across all splits and models, Abortion - Task 1

Argument	F1	Stance	Support	Proportion (in topic)
Gay people should have the same rights as straight people	0.72	PRO	190	0.32
Gay parents can adopt and ensure a happy life for a baby	0.57	PRO	57	0.10
Gay marriages are not normal/against nature	0.53	CON	86	0.14
Religion does not permit gay marriages	0.51	CON	56	0.09
Gay parents cannot raise kids properly	0.51	CON	28	0.05
Gay people have problems and create social issues	0.46	CON	39	0.07
Religion should not be used against gay rights	0.41	PRO	51	0.09
People are born gay	0.40	PRO	91	0.15

Table 12: Average F1 scores, Stance, Support (total counts), and Proportion (in topic) for each argument across all splits and models, Gay Rights - Task 1

Argument	F1	Stance	Support	Proportion (in topic)
Used as a medicine for its positive effects	0.59	PRO	72	0.15
Legalized marijuana can be controlled and regulated by the government	0.55	PRO	141	0.29
Responsible for brain damage	0.55	CON	28	0.06
Prohibition violates human rights	0.53	PRO	93	0.19
If legalized, people will use marijuana and other drugs more	0.52	CON	28	0.06
Damages our bodies	0.40	CON	40	0.08
Highly addictive	0.38	CON	31	0.06
Does not cause any damage to our bodies	0.35	PRO	38	0.08
Causes crime	0.28	CON	17	0.03

Table 13: Average F1 scores, Stance, Support (total counts), and Proportion (in topic) for each argument across all splits and models, Marijuana - Task 1

Argument	F1	Stance	Support	Proportion (in topic)
Healthcare bill is a failure	0.62	CON	25	0.04
Better healthcare	0.59	PRO	27	0.05
Better than the republican candidates	0.51	PRO	69	0.12
Wars are still ongoing	0.51	CON	26	0.05
Created more jobs	0.47	PRO	15	0.03
Destroyed our economy	0.44	CON	74	0.13
Ending the wars	0.43	PRO	30	0.05
Fixed the economy	0.42	PRO	62	0.11
Unemployment rate is high	0.41	CON	14	0.02
Executed effective foreign policies	0.40	PRO	25	0.04
Not eligible as a leader	0.37	CON	56	0.10
Has qualities of a good leader	0.36	PRO	47	0.08
We have better Republicans than Obama	0.26	CON	19	0.03
Ineffective foreign policies	0.26	CON	13	0.02
Makes good decisions/policies	0.30	PRO	35	0.06
Poor decision-maker	0.16	CON	30	0.05

Table 14: Average F1 scores, Stance, Support (total counts), and Proportion (in topic) for each argument across all splits and models, Obama - Task 1

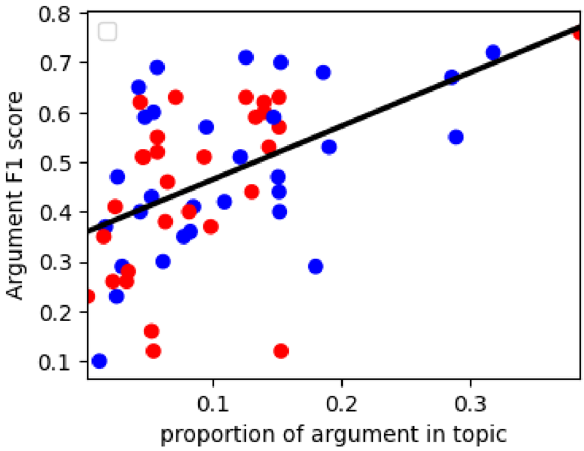


Figure 3: Proportion of each argument within its topic as related to F1 scores (blue = PRO arguments, red = CON arguments)

OLS Regression Results					
Dep. Variable:	y		**R-squared:**	0.262	
Model:	OLS		**Adj. R-squared:**	0.249	
Method:	Least Squares		**F-statistic:**	20.20	
Prob (F-statistic):	3.47e-05		**Log-Likelihood:**	32.081	
No. Observations:	59		**AIC:**	-60.16	
Df Residuals:	57		**BIC:**	-56.01	
Df Model:	1		**Covariance Type:**	nonrobust	
Variable	**coef**	**std err**	**t**	**P>|t|**	**[0.025, 0.975]**
const	0.3569	0.031	11.647	0.000	[0.296, 0.418]
x1	1.0758	0.239	4.494	0.000	[0.596, 1.555]
Omnibus:	2.196		**Durbin-Watson:**	1.130	
Prob(Omnibus):	0.334		**Jarque-Bera (JB):**	1.698	
Skew:	-0.414		**Prob(JB):**	0.428	
Kurtosis:	3.071		**Cond. No.:**	13.0	

Table 15: OLS Regression Analysis